



Current Trends in OMICS

Volume 1 Issue 1, Spring 2021

Journal DOI: <https://doi.org/10/32350/cto>

Homepage: <https://journals.umt.edu.pk/index.php/CTO/index>

Article:

SeqDown: An Efficient Sequence Retrieval Software and Comparative Sequence Retrieval Analysis

Author(s):

Waqar Hanif^{1,3}, Hijab Fatima³, Muhammad Qasim¹, Rana Muhammad Atif², Muhammad Rizwan Javed^{1*}

Affiliations:

¹Department of Bioinformatics and Biotechnology, Government College University Faisalabad (GCUF), Faisalabad, Pakistan

²Department of Plant Breeding and Genetics, University of Agriculture Faisalabad, Faisalabad, Pakistan

³Research Center for Modelling and Simulation, National University of Science and Technology (NUST), Islamabad, Pakistan

[Journal QR](#)



[Article QR](#)



Waqar Hanif

Article History:

Received: June 8, 2021

Revised: June 21, 2021

Accepted: July 19, 2021

Available Online: August 2, 2021

Citation:

Hanif W, Fatima H, Qasim M, Atif RM, Javed MR. SeqDown: an efficient sequence retrieval software and comparative sequence retrieval analysis. *Curr Trend OMICS*. 2021;1(1):18–29.

[Crossref](#)

Copyright
Information



This article is open access and is distributed under the terms of Creative Commons Attribution 4.0 International License



A publication of the

Department of Knowledge and Research Support Services University of
Management and Technology, Lahore, Pakistan

*Corresponding Author: rizwan@gcuf.edu.pk; mrizwanjaved@gmail.com

SeqDown: An Efficient Sequence Retrieval Software and Comparative Sequence Retrieval Analysis

Waqar Hanif^{1,3}, Hijab Fatima³, Muhammad Qasim¹, Rana Muhammad Atif², Muhammad Rizwan Javed^{1*}

¹Department of Bioinformatics and Biotechnology, Government College University Faisalabad (GCUF), Faisalabad, Pakistan

²Department of Plant Breeding and Genetics, University of Agriculture Faisalabad, Faisalabad, Pakistan

³Research Center for Modelling and Simulation, National University of Science and Technology (NUST), Islamabad, Pakistan

Abstract

For any sequence analysis procedure, a single or multiple sequence must be retrieved, stored and organized. Among the most common public databases used for biological sequence retrieval is GenBank, which is a comprehensive public database of nucleotide sequences. However, as the length of the sequence to be retrieved increases such as a chromosome, entire genome, or a scaffold; the elapsed time to download the file is elongated due to slower bandwidth to download/retrieve the sequence.^[8] In most cases of sequence analysis, the researcher requires messenger RNA (mRNA), RNA, DNA, and the protein sequences of the sequence of interest to work with, which consumes a substantial amount of the researcher's time dedicated to find and retrieve the sequence files. Access to GenBank through JAVA HTTPS protocols was established to request and receive the sequence files associated with the input accessions. SeqDown was shown to be very efficient in terms of the retrieval time of the sequences as compared to the other internet browsers and was found to be 15.27% faster than Mozilla Firefox. SeqDown also provides the feature to retrieve coding DNA sequences and protein sequences present in a single chromosome. Sequence retrieval from most biological databases doesn't show the proper naming of their files and the user has to deal with the redundantly named sequence files. This leads to an incorrect and time consuming analysis and this problem can be solved with SeqDown.

Keywords: bioinformatics, biological database, genbank, seqdown, sequence retrieval

*Corresponding Author: rizwan@gcuf.edu.pk; mrizwanjaved@gmail.com

Introduction

Fundamentally, for any sequence analysis procedure a single or multiple sequences must be retrieved, stored, and organized before the actual execution of the sequence analysis [1] methodologies. One of the most common public repositories or databases for submission and retrieval of biological macromolecular sequences is GenBank [2]. GenBank is a comprehensive public database of nucleotide sequences which is built, managed, and distributed by the National Center for Biotechnology Information (NCBI) [3]. It also caters to the translated protein sequences of these archived nucleotide sequences. GenBank synchronizes its data with European Molecular Biology Laboratory European Nucleotide Archive (EMBL-ENA) and DNA Data Bank of Japan (DDBJ) [4, 5] daily to form a collaborative and synchronized data environment where each aforementioned database has the same up-to-date data and information, acting as partners in the International Nucleotide Sequence Database Collaboration (INSDC). [6] Each GenBank record that constitutes a sequence, its related annotation and other information is provided with a unique identifier called an accession number that is shared among all three collaborating participants of the INSDC [1]. The accession number is kept the same throughout the life of the GenBank record however changes to the sequence data are tracked through an integer incremental extension of the accession number. To retrieve the sequence of interest, the user has to visit a particular GenBank webpage associated with the sequence of interest and retrieve its sequence and annotation (if required), the sequence is retrieved through the 'Portal' that is accessible on each page of the GenBank record. Multiple file format options are available for downloading, however, in most cases FASTA [7] format file is retrieved through a specific portal which downloads the respective sequence from the respective GenBank record's record of interest. However, the downloaded file always has the same name for any sequence that is retrieved which is "*sequence.fasta*", which makes it quite difficult for the researcher. The user has to keep track of the sequences that were retrieved because it is a good practice in bioinformatics analysis to name the files according to their description. As the length of the sequence to be retrieved becomes longer such as a chromosome, entire genome or a scaffold, the elapsed time to download the file through Portal (the default option on GenBank to download the sequence) gets even elongated due to slower bandwidth to download/retrieve the sequence. [8] The bandwidth varies from browser to browser, some of which seem to work efficiently in retrieving the file quickly.

In most cases, during sequence analysis, the researcher requires messenger RNA, RNA, DNA, or even protein sequences of the same sequence of interest to

work with, which consumes a substantial time of the researcher in finding and retrieving the aforementioned sequence files. [2] In some cases, it is a necessity to retrieve all of the protein sequences or CDSs that are constituted by a single chromosome which can become a cumbersome task for non-advanced database users. [9] There's no proper methodology available to do it through a single platform or through one-click functionality, for example, retrieval of mRNA, DNA, RNA, and protein sequence of a single gene by just providing one single accession of either of the aforementioned required/desired sequences.

In most cases, sequences-of-interest are retrieved repeatedly by visiting the GenBank record webpages. There are other tools available that somewhat deal with these problems, but they aren't locally installed tools and have to be utilized through web servers. The sequences however, can be retrieved from these online databases but it is very cumbersome and complex and there is no other standalone platform previously available for this purpose. SeqDown categorically deals with the aforementioned problems and provides a novel approach to retrieve/download the sequences efficiently and effectively especially when it comes to downloading the sequences in the quickest way possible. Hence, to solve the aforementioned problems: retrieve mRNA, CDS, Flat File, complete sequence just by providing a single accession, have the software installed locally on the computer (Windows, Linux, macOS), download the sequences in an efficiently quickest way by utilizing the network bandwidth at its full potential and saving those downloaded sequences into their descriptively named files.

Materials and Methods

SeqDown works with both GenBank, GenPept accession numbers and RefSeq [10] accessions numbers. SeqDown is developed using the JAVA [11] programming language on top of the JAVAFX [12] framework platform.

Input

SeqDown accepts GenBank and RefSeq nucleotide or protein (GenPept) accession numbers with or without an incremental integer, that is, for example the accession number NM_002299 [13] can be input as well as NM_002299.x (x is its incremental integer). Utilizing the former as input will automatically yield a result of downloading the latest entry for the provided GenBank/RefSeq accession, whereas utilizing the latter will yield a result of downloading only the exact entry for the provided GenBank/RefSeq accession. The algorithm that SeqDown is based upon is divided into 2 modules depending upon the input that is, whether the input is a protein accession or a nucleotide accession, therefore executing only the respective and required functions.

Figure 1. A Screenshot of the Seqdown Tool. The Accession ID is Entered and Download is Pressed for a Single Click Retrieval of the Sequence

SeqDown

Nucleotide Accession: Enter Nucleotide Accession Number Here!

OR

Protein Accession: NP_002290.2

Download

Retrieval Status...

Exit

When an Accession Number is Entered, the program will retrieve the following:

- Complete Sequence
- CDS (in case coding sequence of a gene/protein is available)
- Complete Flat File
- Protein Sequence

Department of Bioinformatics and Biotechnology
Government College University Faisalabad

Nucleotide Sequence Accession as Input and its Retrieval

When a nucleotide sequence accession number is provided as an input, the code modules that are involved in the retrieval, file creation, directory handling, file handling, and saving/storage of the retrieved sequence(s) are executed. First and foremost, a connection to GenBank/RefSeq is made to test a stable connection for data handling and data retrieval. Once the connection has been made, the input nucleotide sequence accession number is utilized to retrieve its protein sequence, coding DNA sequence (CDS, if it exists), its complete nucleotide sequence, and complete GenBank file (which contains various bibliographic information and annotation related to the input accession number) one by one which is then stored on the computer. However, it should be noted that chromosome accession number as input yields is a result of retrieval of all the CDSs and protein sequences that are constituted within the chromosome.

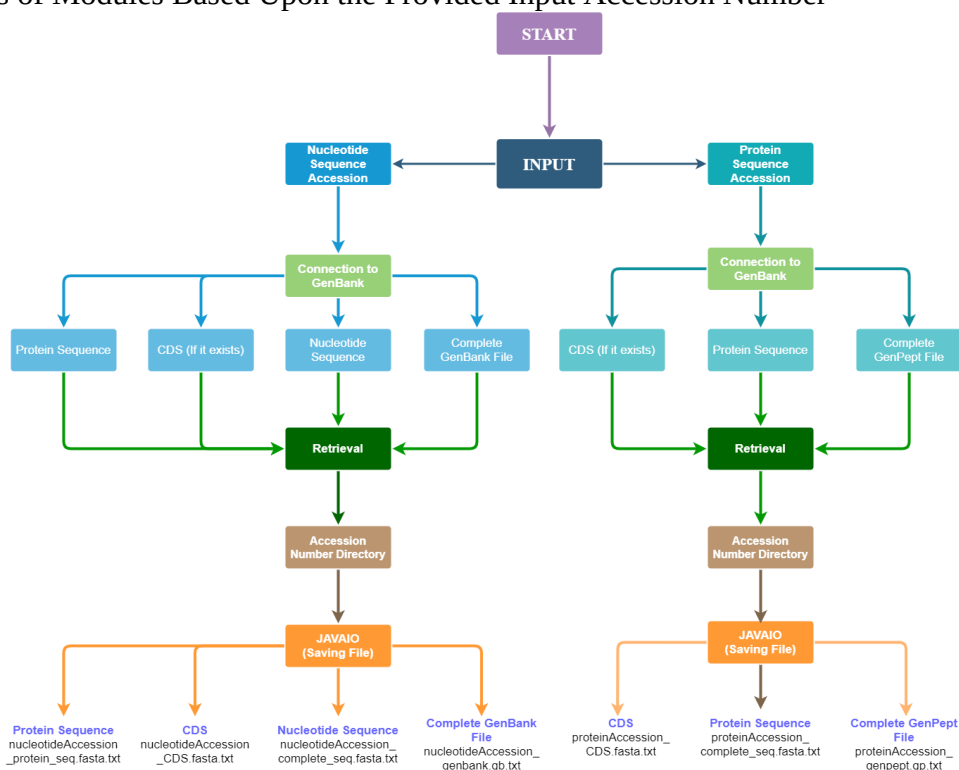
Protein Sequence Accession as Input and its Retrieval

When a protein sequence accession number is provided as an input, the code modules that are involved in the retrieval, file creation, directory handling, file handling and saving of the retrieved sequence(s) are executed. First and foremost, a connection to GenBank/RefSeq is made to test a stable connection for data handling and data retrieval. Once the connection has been made, the input protein sequence accession number is utilized to retrieve its protein sequence, coding DNA sequence (CDS, if it exists) and complete GenPept (a database that archives the translated protein sequences of the coding regions from GenBank) file (which contains various bibliographic information and annotation related to the input protein accession number) one by one, which is then stored on the computer.

File Handling and Storage

Once the files have been retrieved from GenBank in case of a nucleotide accession number as an input, the retrieved files are iteratively passed to *JAVAI/O* (JAVA Input/Output) for the creation of a directory on the Desktop (in case of Microsoft Windows [14] or Home (in case of Linux [15] and Apple macOS [16] with the provided input as its name. Retrieved *protein* sequence, *CDS*, complete *nucleotide* sequence and the complete *GenBank* files are saved into the newly created directory with *nucleotide_Accession_protein_seq.fasta.txt*, *nucleotideAccession_CDS.fasta.txt*, *nucleotideAccession_complete_seq.fasta.txt*, *nucleotideAccession_FLAT_file.fasta.txt* as names respectively. The files then can be utilized for further analysis in any of the tools, software or other methodologies.

In the case of a protein accession number as an input, the retrieved files are iteratively passed to *JAVAI/O* (JAVA Input/Output) for the creation of a directory on the Desktop (in case of Windows) or Home (in case of Linux and macOS) with the provided input as its name. Retrieved *protein* sequence, *CDS* and the complete *GenPept* files are saved into the newly created directory with *ProteinAccession_CDS.fasta.txt*, *protein_Accession_complete_seq.fasta.txt*, *protein_GenPept_file.fasta.txt* as names respectively. The .txt extension is utilized to make it easier for everyone to open and browse the files in the default .txt file opening programs of the operating systems, such as Notepad on Windows and TextEdit on macOS. However, in some cases FASTA or GenBank extension is required, therefore the .txt extension can be removed which will make the extension as FASTA or GenBank respectively. The files then can be utilized for further analysis in any of the tools, software or other methodologies. Each retrieval execution creates a folder corresponding to the input accession number as its name. The flow of the software can be comprehended through its flowchart.

Figure 2. Flowchart of the Seqdown Algorithm – Seqdown Executes Two Different Sets of Modules Based Upon the Provided Input Accession Number

Results

Sequence Retrieval

To test the software, we used the accession number of a nucleotide GenBank record *NM_002299.4* that is assigned to *Homo sapiens lactase (LCT), mRNA*. SeqDown retrieved the CDS, complete nucleotide sequence, protein sequence and the complete GenBank file of the GenBank record *NM_002299.4* and saved it into a folder on the desktop in a folder named, *NM_002299.4*, the names that were assigned to the retrieved sequences and their files are displayed in Table 1 (A). We also tested the software using the accession number of a protein GenBank record, *NP_002290.2* [17] that is assigned to *lactase-phlorizin hydrolase preproteins [Homo sapiens]*. SeqDown retrieved the CDS, complete nucleotide sequence, protein sequence and the complete GenBank file of the GenBank record *NP_002290.2* and saved it into a folder on the desktop in a folder named, *NP_002290.2*, the names that were assigned to the retrieved sequences and their files are displayed in the Table 1 (B).

Figure 3. A Screenshot of the SeqDown Tool when the Sequence is Retrieved. A Folder is Created on the Desktop with the Accession ID as the File Name

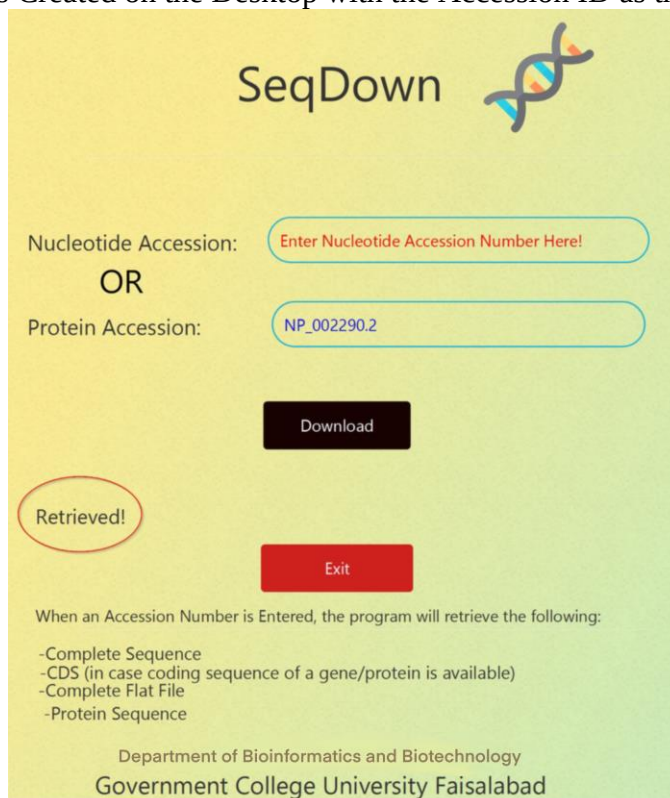


Table 1. Sequence Retrieval of Provided Input Accession Number Using SeqDown
(A) Nucleotide Accession Number

Sequence/File Type	File (Name)
Nucleotide	NM_002299.4_complete_seq.fasta.txt
Protein	NM_002299.4_protein_seq.fasta.txt
CDS	NM_002299.4_CDS.fasta.txt
GenBank (File)	NM_002299.4_GenBank.gb.txt

(B) Protein Accession Number

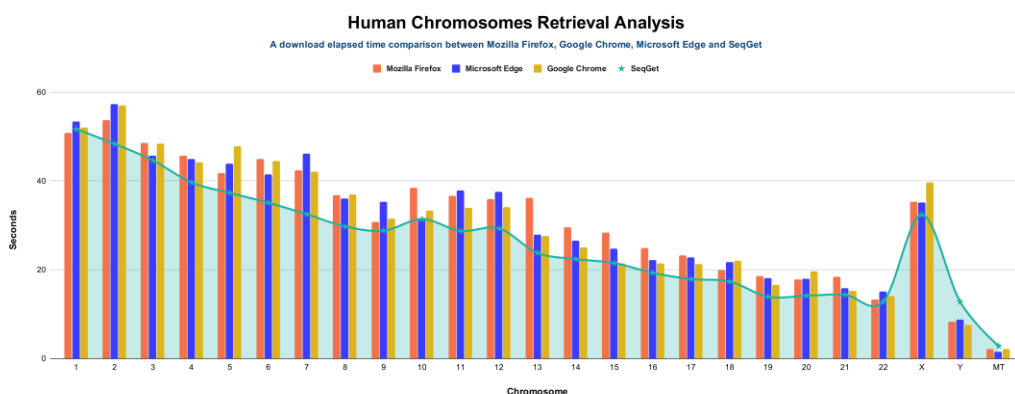
Protein	NP_002290.2_complete_seq.fasta.txt
CDS	NP_002290.2_CDS.fasta.txt
GenPept	NP_002290.2_GenPept.gp.txt

–(A) The provided input was a nucleotide accession number NM_002299.4 which yielded the retrieval of four different files including nucleotide, protein, CDS, and

GenBank file associated with the accession. (B) The provided input was a protein accession number NP_002290.2 which yielded the retrieval of three different files including protein, CDS, and GenPept file associated with the accession

Comparative Sequence Retrieval Analysis

Figure 4. Human Chromosomes Retrieval Analysis



—This analysis was done using the most commonly used internet browsers such as Mozilla Firefox, Google Chrome, Microsoft Edge and SeqDown. It was shown that SeqDown was able to retrieve all 23 human chromosome sequences including the sex chromosomes X, Y, and mitochondrial DNA collectively sooner than the internet browsers used in this study

To determine the efficiency and speed of the retrieval of SeqDown, we compared the time it took for SeqDown against the most commonly used internet browsers such as Mozilla Firefox, Google Chrome, and Microsoft Edge to retrieve the nucleotide sequences of all the 23 Homo sapiens chromosomes [18] (22 chromosomes including X & Y sex chromosome) and the mitochondrial chromosome of the Homo sapiens. We used the RefSeq accession number of each chromosome and only downloaded the nucleotide sequence through SeqDown, Mozilla Firefox, Google Chrome, and Microsoft Edge. Once the entire set of Homo sapiens chromosome through each internet browser and SeqDown was downloaded, the total time was calculated. It became quite evident that SeqDown was able to retrieve the chromosomal sequences sooner as compared to the internet browsers. It also became evident that internet browsers didn't keep a stable bandwidth to retrieve/download the sequences, rather the bandwidth fluctuated frequently which slowed down the downloading and therefore, increased the time elapsed to retrieve the sequences. However, SeqDown on the other hand was shown to keep a stable bandwidth when retrieving the sequences which increased the

downloading and therefore, decreased the time elapsed to retrieve the sequences. All the Homo sapiens chromosomes were retrieved in a total time of 662.69, 759.9, 769.44, and 782.21 seconds using SeqDown, Google Chrome, Microsoft Edge and Mozilla Firefox respectively. The comparison according to the retrieval time of each chromosome is shown in Figure 4. SeqDown was found to be 12.8% faster than Google Chrome, 13.87% faster than Microsoft Edge and 15.27% faster than Mozilla Firefox. Inputting the accession number of the 1st and only chromosome [19] of *severe acute respiratory syndrome coronavirus 2* allowed us to retrieve all the CDSs and protein sequences that were constituted within the chromosome, a total of 12 protein and coding DNA sequences were retrieved which were later verified from NCBI Genomes [20].

Discussion and Conclusion

The traditional way of retrieving the sequences through the databases involves the webserver of those particular databases which is cumbersome and a complex method. Also, there are no standalone platforms previously available for the purpose. Furthermore, it is generally considered a good practice in bioinformatics analysis to name and organize the project files according to their respective constituents. General sequence retrieval from most biological databases ignores this and the user has to deal with the redundantly named sequence files which lead to incorrect and time-consuming analysis. These aforementioned problems can be effectively solved with the usage of SeqDown. SeqDown retrieves mRNA, CDS, Flat File, complete sequence just with a single click by providing an accession ID to an easily downloadable tool on a computer (Windows, Linux, macOS) unlike the other databases where the user needs to retrieve all the data one by one. This retrieval of the sequences in a single step could be done in a systematic way as well while using biopython but that includes a command line which is not user friendly. Whereas the biggest benefit of SeqDown tool is for those who are the beginners in this field and find the sequences retrieval from the databases a complex and time consuming process.

Availability and Requirements

- *Project name:* SeqDown
- *Project home page:* <https://sourceforge.net/projects/seqdown/files/>
- *Operating system(s):* Windows, macOS, Linux
- *Programming language:* JAVA
- *Other requirements:* At least 185 megabytes of storage, 100 megabytes of free random-access memory, and an internet connection.
- *License:* Free

- Any restrictions to use by non-academics: None

Consent for Publication

All authors provide their consent for this publication.

Conflict of Interest

None

Acknowledgments

All authors acknowledge the platform provided by the Research Centre of Modelling and Simulation, National University of Science and Technology and Department of Bioinformatics and Biotechnology, Government College University Faisalabad.

Funding

None

References

- [1] Long K, Cai L, He L. DNA Sequencing Data Analysis. In *Computational Systems Biology 2018* (pp. 1-13). Humana Press, New York, NY.
- [2] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, Sayers EW. GenBank. *Nucleic Acid Res.* 2018;46(D1):D41-7.
- [3] Coordinators NR. Database resources of the National Center for Biotechnology Information. *Nucleic Acid Res.* 2017;46(D1):D8-D13. <https://doi.org/10.1093/nar/gkx1095>
- [4] Leinonen R, Akhtar R, Birney E, et al. The European nucleotide archive. *Nucleic Acid Res.* 2010;39(suppl_1):D28-31. <https://doi.org/10.1093/nar/gkq967>
- [5] Kodama Y, Mashima J, Kosuge T, Ogasawara O. DDBJ update: the Genomic Expression Archive (GEA) for functional genomics data. *Nucleic Acid Res.* 2019;47(D1):D69-73. <https://doi.org/10.1093/nar/gky1002>
- [6] Karsch-Mizrachi I, Takagi T, Cochrane G, International Nucleotide Sequence Database Collaboration. The international nucleotide sequence database collaboration. *Nucleic Acid Res.* 2018;46(D1):D48-51. <https://doi.org/10.1093/nar/gkx1097>
- [7] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proceedings of the Nat Acad Sci.* 1988;85(8):2444-8. <https://doi.org/10.1073/pnas.85.8.2444>

- [8] ASPERA connect [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020 06 06]. Available from: <https://www.ncbi.nlm.nih.gov/public/>
- [9] Doležel J, Vrána J, Šafář J, Bartoš J, Kubaláková M, Šimková H. Chromosomes in the flow to simplify genome analysis. *Funct Integrative Genomics*. 2012;12(3):397-416. <https://doi.org/10.1007/s10142-012-0293-0>
- [10] O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acid Res*. 2016;44(D1):D733-45. <https://doi.org/10.1093/nar/gkv1189>
- [11] Arnold K, Gosling J, Holmes D. *The Java Programming Language*. Addison Wesley Professional; 2005.
- [12] OpenJDK: OpenJFX Project. <https://openjdk.java.net/projects/openjfx/>
- [13] Homo sapiens lactase. 2020. <https://www.ncbi.nlm.nih.gov/gene?Cmd=DetailsSearch&Term=3938>
- [14] Bott E, Stinson C. *Windows 10 inside out*. Microsoft Press; 2019.
- [15] Sobell MG. *A practical guide to Ubuntu Linux*. Pearson Education; 2015.
- [16] Apple Inc. macOS High Sierra [Internet]. 2019. <https://www.apple.com>
- [17] NCBI. Lactase-phlorizin hydrolase preproprotein [Homo sapiens]. https://www.ncbi.nlm.nih.gov/protein/NP_002290.2
- [18] NCBI. Homo sapiens (Human). Genome. <https://www.ncbi.nlm.nih.gov/genome/?term=homo+sapiens>
- [19] Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome. 2020. <https://www.ncbi.nlm.nih.gov/nuccore/1798174254>
- [20] Severe acute respiratory syndrome coronavirus 2 (ID 86693) - Genome - NCBI. <https://www.ncbi.nlm.nih.gov/genome/86693>

Supportive Material

The time-series data for the sequence retrieval can be accessed through the Source Forge project page.